RECOGNITION OF HUMAN ACTIVITIES USING SPACE DEPENDENT SWITCHED DYNAMICAL MODELS

Jacinto C. Nascimento [a] *Mário A. T. Figueiredo* [b] *Jorge S. Marques* [a]

Instituto de Sistemas e Robótica [a] Instituto de Telecomunicações [b]
Instituto Superior Técnico
1049-001 Lisboa,
Portugal

ABSTRACT

This paper describes a new algorithm for the recognition of human activities. These activities are modelled using banks of switched dynamical models, each of which is tailored to a specific motion regime. Furthermore, it is assumed that model switching happens according to a space-dependent Markov chain, i.e., some transitions are more probable in specific regions of the image. Space dependence allows the model to represent the interaction between the person and static elements of the scene. The paper describes learning algorithms for space-dependent switched dynamical models and presents experimental results with synthetic and real data.

1. INTRODUCTION

The analysis of human activities is a key step in several applications of video processing, such as human-machine interaction, video surveillance, and smart rooms [1]. The problem has been addressed using different representations of the human body. Some works represent the human body as a set of segments and try to characterize the position of each segment during the video sequence [2, 3]; activities (e.g., walking) are then characterized from physiological parameters. However, estimation of articulated models of the full human body remains a difficult task which has not been solved in a robust way. Many authors try to represent the human body and its evolution during the video sequence without segmenting it; the body is either represented by a deformable contour [4], by templates [5], or simply by a single blob [6]. In the latter case, the human activity is characterized by the evolution (trajectory) of a reference point (e.g., center of mass) in the video sequence. Activity recognition from trajectories has been addressed using statistical models, e.g., hidden Markov models (HMMs), coupled hidden Markov models (CHMMs) [6], or Bayesian networks [7].

In this paper, we represent the human activity by the trajectory of the centroid, as in other previous publications, since it is a robust and sufficient feature for many applications. The evolution of the centroid in each activity is then represented using a dynamical model. Unfortunately, a single model is not enough to cope with complex trajectories (e.g., a person entering a shop). Therefore, multiple models are used, and a switching mechanism is considered. Furthermore, the switching probabilities should be space dependent in order to be able to represent the interaction between a person and the scene (e.g., people often change direction close to the store entrance; see Fig. 1). Therefore, each activity is characterized by a space-dependent switched dynamical model (SDSDM).

In Section 2, we formally define the SDSDM. Then, in Section 3, we derive the learning method used to estimate the model parameters from the video stream. Activity recognition is addressed in Section 4. Section 5 presents experimental results and Section 6 concludes the paper.

Fig. 1. Shopping mall with the area of interest.

2. ACTIVITY MODEL

We assume that the human activities of interest can be recognized from the trajectory of each person on the image plane, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, with $\mathbf{x}_t \in I\!R^2$. Furthermore, we assume that this trajectory is the output of a bank of switched dynamical systems of the form

$$\mathbf{x}_t = \mathbf{x}_{t-1} + T_{k_t} + w_t, \tag{1}$$

where $k_t \in \{1, ..., m\}$ is the label of the active model at time instant t, T_{k_t} is a (model-dependent) displacement vector, and the $w_t \sim \mathcal{N}(0, Q_{k_t})$ are independent Gaussian random variables, with (also model-dependent) covariances Q_{k_t}.

We assume that $k = (k_1, k_2, \ldots, k_n)$ is a Markov sequence with a space-dependent transition matrix, i.e., the transition probabilities depend on the location of the person, i.e., on $\mathbf{x}_t$ (see Fig.

This work was supported by the (Portuguese) Foundation for Science and Technology (FCT), under project LTT and by the EU, under project CAVIAR (IST-2001-37540).

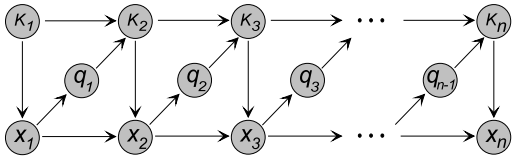


Fig. 2. Architecture of the proposed SDSDM: k_t - label model (hidden variables); q_t binary variable, $\mathbf{x}_t$ state (observations).

2). For the sake of simplicity, the image plane is split into a set of disjoint regions R_i, for $i = 1, \ldots, d$, and a different transition matrix B_i is assigned to each region. Fig. 1 shows an example in which two regions are considered: R_0, which accounts for the corridor, and R_1, corresponding to the shop entrance in which most model transitions occur. This idea can be formalized as follows. Let q_t be a discrete variable defined by the following deterministic relation:

$$q_t = p \qquad \Leftrightarrow \qquad \mathbf{x}_t \in R_p. \tag{2}$$

Then, we will assume that

$$P(k_t = j | k_{t-1} = i, q_{t-1} = p) = B_p(i, j). \tag{3}$$

The relationship among the variables $\mathbf{x}, k, q$ is represented by the graphical model shown in Fig. 2, where $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ is the observed trajectory, $q = (q_1, ..., q_n)$ is simply a deterministic function of $\mathbf{x}$, and $k = (k_1, k_2, \ldots, k_n)$ is a hidden sequence. This model depends on a set of parameters, denoted collectively as $\theta = (T_1, Q_1, \ldots, T_m, Q_m, B_1, \ldots, B_d)$, which have to be estimated from data.

3. LEARNING ALGORITHM

Due to the conditional dependence of k_t, not only on k_{t-1}, but also on $\mathbf{x}_{t-1}$ (via q_{t-1}), the standard *Baum-Welch algorithm* (BWA) [8] is not directly applicable to learn the parameters of this model. In this section, we present a modified BWA, tailored to the model described in the previous section. Just as the BWA is simply an instance of the expectation-maximization (EM) algorithm [9] to estimate the parameters of a standard HMM, our algorithm is EM applied to the model described in the previous section.

The complete log-likelihood function $L(\theta) = \log p(\mathbf{x}, k \mid \theta)$, assuming that all the variables were observed, is given by

$$L(\theta) = \sum_{t=2}^{n} \big\{ \log p(\mathbf{x}_t | \mathbf{x}_{t-1}, k_t) + \log B_{q_{t-1}}(k_{t-1}, k_t) \big\} + L_1, \tag{4}$$

where $L_1 = \log p(\mathbf{x}_1, k_1)$, which we admit known to simplify the notation. Also to simplify the notation, we omit the explicit dependency on θ of the right hand side of (4), and of most other equations in this section. The EM algorithm produces a sequence of parameter estimates $\hat{\theta}^1, \ldots, \hat{\theta}^s, \theta^{s+1}, \ldots$ by iteratively maximizing an auxiliary function,

$$\hat{\theta}^{s+1} = \arg\max_{\theta} Q(\theta; \hat{\theta}^s), \tag{5}$$

where $Q(\theta; \hat{\theta}^s)$ is the conditional expectation of the complete log-likelihood, with respect to the hidden sequence k, given the current parameter estimates $\hat{\theta}^s = (\hat{T}_1^s, \hat{Q}_1^s, \ldots, \hat{T}_m^s, \hat{Q}_m^s, \hat{B}_1^s, \ldots, \hat{B}_d^s)$, and the observed sequence x; that is,

$$Q(\theta; \hat{\theta}^s) = E_k \Big\{ \log p(\mathbf{x}, k \mid \theta) \mid \mathbf{x}, \hat{\theta}^s \Big\}. \tag{6}$$

The conditional probability of the hidden state sequence k, given $\mathbf{x}$ and $\hat{\theta}^s$, necessary to obtain $Q(\theta; \hat{\theta}^s)$, is computed in the E-step. This is carried out, as in the BWA, in two recursions, forward and backward, leading to $p(k_t | \mathbf{x}_1^t)$, $p(k_{t-1} | \mathbf{x})$ and $P(k_{t-1}, k_t | \mathbf{x})$, with the notation $\mathbf{x}_1^t = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$.

Forward recursion: the *prediction step* is given by

$$P(k_t = j | \mathbf{x}_1^{t-1}) = \sum_{i=1}^{m} \hat{B}_{q_{t-1}}^s(i, j) P(k_{t-1} = i | \mathbf{x}_1^{t-1}), \tag{7}$$

where q_{t-1} is defined in (2). The *filtering step* is given by

$$P(k_t = j | \mathbf{x}_1^t) = \frac{P(\mathbf{x}_t | k_t = j, \mathbf{x}_{t-1}) P(k_t = j | \mathbf{x}_1^{t-1})}{\sum_{i=1}^{m} P(\mathbf{x}_t | k_t = j, \mathbf{x}_{t-1}) P(k_t = j | \mathbf{x}_1^{t-1})}. \tag{8}$$

Backward recursion: Here the goal is to obtain $P(k_{t-1}, k_t | \mathbf{x})$ and $P(k_t | \mathbf{x})$, which is done following

$$\begin{aligned} P(k_{t-1} = i, k_t = j | \mathbf{x}) &= P(k_{t-1} = i | k_t = j, \mathbf{x}) P(k_t = j | \mathbf{x}) \\ &= \hat{B}_{q_{t-1}}^s(i, j) \frac{P(k_{t-1} = i | \mathbf{x}_1^{t-1}) P(k_t = j | \mathbf{x})}{P(k_t = j \mid \mathbf{x}_1^{t-1})}. \end{aligned} \tag{9}$$

and

$$\begin{aligned} P(k_{t-1} = i | \mathbf{x}) &= \sum_{j=1}^{m} P(k_{t-1} = i, k_t = j | \mathbf{x}) \\ &= P(k_{t-1} = i | \mathbf{x}_1^{t-1}) \sum_{j=1}^{m} \frac{\hat{B}_{q_{t-1}}^s(i, j) P(k_t = j | \mathbf{x})}{P(k_t = j | \mathbf{x}_1^{t-1})}. \end{aligned} \tag{10}$$

Recall that all probabilities in (7)–(10) are conditioned on $\hat{\theta}^s$.

Equation (6) can be now be written straightforwardly as

$$\begin{aligned} Q(\theta; \hat{\theta}^s) &= \sum_{t=2}^{n} \sum_{i=1}^{m} P(k_t = i | \mathbf{x}, \hat{\theta}^s) \log p(\mathbf{x}_t | k_t = i, \mathbf{x}_{t-1}, \theta) \\ &\quad + \sum_{t=2}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} P(k_t = i, k_{t-1} = j | \mathbf{x}, \hat{\theta}^s) \log B_{q_{t-1}}(i, j), \end{aligned}$$

leading to

$$\begin{aligned} Q(\theta; \hat{\theta}^s) &= \sum_{t=2}^{n} \sum_{i=1}^{m} w_i^t \Big(-\frac{1}{2} \log \det(Q_i^s) - \frac{1}{2} (\nu_i^t)^T (Q_i^s)^{-1} \nu_i^t \Big) \\ &\quad + \sum_{t=2}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij}^t \log B_{q_{t-1}}(i, j), \end{aligned} \tag{11}$$

where $w_i^t = P(k_t = i | \mathbf{x}, \hat{\theta}^s)$, $w_{ij}^t = P(k_t = i, k_{t-1} = j | \mathbf{x}, \hat{\theta}^s)$ and $\nu_i^t = \mathbf{x}_t - \mathbf{x}_{t-1} - \hat{T}_i^s$ is the prediction error.

Denoting $\mathbf{T} = (T_1, \ldots, T_m)$, $\mathbf{Q} = (Q_1, \ldots, Q_m)$, and $\mathbf{B} = (B_1, ..., B_d)$, it's clear that (11) can be maximized separately with respect to $\mathbf{B}$ and the pair $(\mathbf{T}, \mathbf{Q})$. After straightforward manipulation, we obtain

$$\hat{B}_p^{s+1} = \textit{normalize-rows}(A_p) \tag{12}$$

where *normalize-rows*$(\cdot)$ is an operator that normalizes each row of a matrix with non-negative entries, to guarantee that it is a valid stochastic matrix, and

$$A_p(i,j) = \alpha + \sum_t w_{ij}^t \ \delta_{p,q_t}, \tag{13}$$

with α a small regularization constant, and δ_{p,q_t} the Kronecker delta function, i.e., $\delta_{p,q_t} = 1$, if $q_t = p$, and zero otherwise. The regularization constant α is necessary because we observe a very small number (typically zero or one) of model transitions. The other parameters, $(\mathbf{T}, \mathbf{Q})$, are updated according to standard rules of the BWA, and we refer the reader to [8] for details; notice that (11), with respect to $(\mathbf{T}, \mathbf{Q})$, is formally equivalent to the Q-function in the EM algorithm for estimating a mixture of m Gaussians [9].

Summarizing, the EM algorithm comprises two steps in each iteration: the E-step, which computes the probabilities of the hidden variables, using (7)-(10), and the M-step in which parameter estimates are updated. All these equations are trivially extended to the case where, instead of one sequence, we have a set of observed sequences. For the sake of simplicity, we omit these details.

4. CLASSIFICATION

Let $\mathcal{A} = \{A_1, \ldots, A_l\}$ be the set of activities according to which we want to classify the observed trajectories. A switched dynamical model θ_i, is estimated from the data for each activity A_i, using the EM algorithm described in Section 3.

The classification of a new sequence $\mathbf{x}$ is obtained by the *maximum a posteriori* (MAP) rule

$$j = \arg\max_i \{p(\mathbf{x}|\hat{\theta}_i)p(A_i)\}; \tag{14}$$

in this paper we consider uniform prior probabilities $p(A_i) = 1/l$. To compute (14), we observe that

$$\begin{aligned} p(\mathbf{x} \mid \hat{\theta}_i) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\hat{\theta}_i) \\ &= \prod_{t=2}^{n} p(\mathbf{x}_t \mid \mathbf{x}_1^{t-1}, \hat{\theta}_i)p(\mathbf{x}_1|\hat{\theta}_i); \end{aligned} \tag{15}$$

since each term of $p(\mathbf{x} \mid \hat{\theta}_i)$ is given by

$$p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, \hat{\theta}_i) = \sum_{i=1}^{m} P(\mathbf{x}_t|k_t = i, \mathbf{x}_{t-1}, \hat{\theta}_i)P(k_t = i|\mathbf{x}_1^{t-1}, \hat{\theta}_i),$$

the observed likelihood can be computed from the denominator of (8). Thus, given a new sequence $\mathbf{x}$, we run one half iteration (the forward recursion) of the E-step under all candidate classes $\{A_1, \ldots, A_l\}$. This gives us $\{p(\mathbf{x}|\hat{\theta}_i), i = 1, \ldots, l\}$. The most likely class is then simply

$$\arg\max_i \{p(\mathbf{x}|\hat{\theta}_i)\} \quad i = 1, \ldots, l. \tag{16}$$

5. EXPERIMENTAL RESULTS

This section presents experimental results using synthetic and real data. In the synthetic case, we have considered a bank of SDSDM and performed Monte Carlo tests. We have defined four classes ($l = 4$) which represent typical activities performed by humans in shopping malls: entering shop, leaving the shop, passing, and browsing. Each of the four activities is represented by the corresponding parameter estimates $\hat{\theta}_i$, $i = 1, 2, 3, 4$, obtained using the EM algorithm described in Section 3, from sample trajectories generated according to (1). For all the activities, we assume two dynamical models, i.e., $m = 2$. Several sample trajectories of each activity are shown in Fig. 3. This figure depicts a store entrance and a corridor. The dashed square represents the area of interest, R_1, where state switching is more probable to occur. We have also generated 100 sequences of each activity to be used as test data. All the test sequences were correctly classified (100% accuracy).

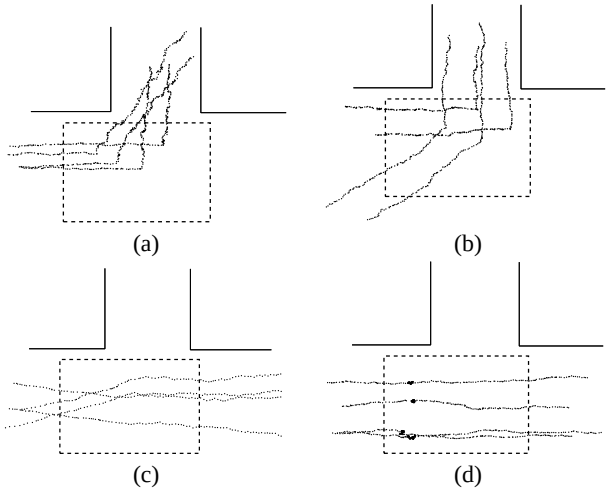


Fig. 3. Several synthetic activities considered: (a) entering, (b) leaving, (c) passing, (d) browsing.

The proposed algorithm was also tested with real data collected in the context of the EU funded project CAVIAR. The data was collected and the ground truth was hand-labelled for 40 video sequences comprising about 90K frames.[1] These sequences include indoor plaza and shopping center observations of individuals and small groups of people. The sequences are labelled with both the tracked persons and also a semantic description of their activities. Fig. 4 shows several real trajectories of the centroid of the bounding box of each person. To obtain the results shown in Table 1, two movies of 5 minutes each were selected: *TwoEnterShop1Front* and *TwoEnterShop2Front*. To test the performance of the algorithm, all the activities presented in the first one were used for training. The seven activities contained in the second movie (three "passings", two "enterings", and two "leavings") were considered as test samples to be classified. Table 1 shows the log-likelihood of each learned model for each activity of the test sequence; we can see that all the trajectories were correctly classified, that is, they exhibit the highest log-likelihood for the correct class.

[1]The ground truth labelled video sequences is provided at `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`.

Input Activities	Test trajectories						
	P_1	P_2	P_3	E_1	E_2	L_1	L_2
P	**-519.1**	**-686.5**	**-546.8**	-1542.4	-1559.0	-1737.1	-1752.0
E	-605.1	-690.6	-567.7	**-450.1**	**-361.1**	-651.2	-632.3
L	-1175.3	-1261.8	-1127.2	-921.8	-814.3	**-424.2**	**-332.7**

Table 1. Log-likelihood classification of real activities: E- entering, L-leaving, P- passing.

6. CONCLUDING REMARKS

In this paper we have proposed and tested an algorithm for modelling and recognizing human activities in a constrained environment. The proposed approach uses a switched dynamical model in which model switching depends on the space variable. It is demonstrated that the proposed model provides good results with synthetic and real data obtained in a commercial center. The proposed method is able to effectively recognize instances of learned activities. The activities studied herein can be interpreted as atomic ones, in the sense that they are simple events. We plan to conduct more extensive tests and to represent complex behaviors as concatenations of the activities studied in the paper.

Acknowledgement: We would like to thank Prof. José Santos Victor of ISR and the members of CAVIAR project, for providing video data of human activities with the ground truth information.

7. REFERENCES

[1] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 780–785, July 1997.

[2] O. Masoud and N.P. Papanikolopoulos, "A method for human action recognition," *Image and Vision Computing*, vol. 21, no. 8, pp. 729–743, August 2003.

[3] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 232–247, February 1999.

[4] N. Johnson and D. Hogg, "Representation and synthesis of behaviour using Gaussian mixtures," *Image and Vision Computing*, vol. 20, no. 12, pp. 889–894, 2002.

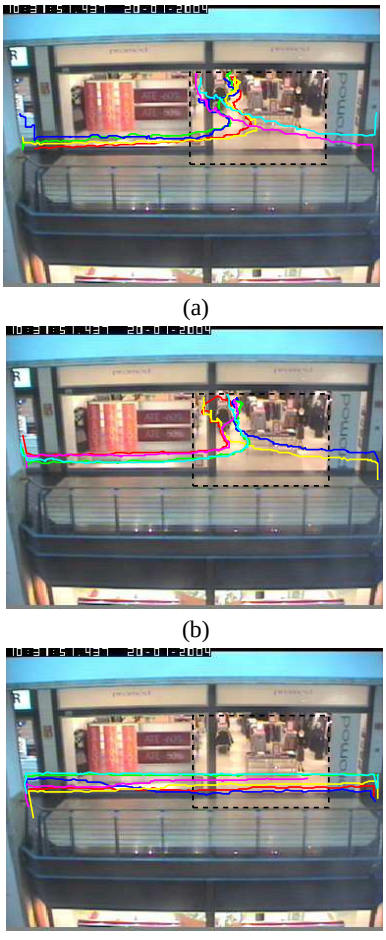
[5] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, March 2001.

[6] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, no. 8, pp. 831–843, August 2000.

[7] S. Hongeng and R. Nevatia, "Multi-agent event recognition," 2001, vol. 2, pp. 84–91, In Proc. of the 8 th IEEE International Conference on Computer Vision (ICCV'01).

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[9] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm.," *Jour. Royal Statist. Soc. (B)*, vol. 39, pp. 1–38, 1977.



(a) (b) (c)

Fig. 4. Several synthetic activities are considered: (a) entering, (b) leaving, (c) passing.